PERGAMON

Neural Networks 13 (2000) 589–596

Neural Networks

www.elsevier.com/locate/neunet

Blind source separation in the presence of weak sources

J.-P. Nadal[a], E. Korutcheva[b,*], F. Aires[a,2]

[a]*Laboratoire de Physique Statistique de l'ENS[3] Ecole Normale Supérieure 24, rue Lhomond-75231 Paris Cedex 05, France*
[b]*Departamento de Física Fundamental, Universidad Nacional de Educación a Distancia (UNED), c/Senda del Rey No 9-28080 Madrid, Spain*

Received 23 June 1999; accepted 26 April 2000

Abstract

We investigate the information processing of a linear mixture of independent sources of different magnitudes. In particular we consider the case where a number m of the sources can be considered as "strong" as compared to the other ones, the "weak" sources. We find that it is preferable to perform blind source separation in the space spanned by the strong sources, and that this can be easily done by first projecting the signal onto the m largest principal components. We illustrate the analytical results with numerical simulations. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Independent component analysis; Blind source separation; Infomax

1. Introduction

During the recent years many studies have been devoted to the study of Blind Source Separation (BSS) and more generally to that of Independent Component Analysis (ICA) (see, e.g. Amari & Cardoso, 1997; Cardoso, 1989; Comon, 1994; Herault, Jutten & Ans, 1985). Within the standard framework one assumes a multidimensional measured signal to result from a linear mixture of statistically independent components, or "sources". In most cases one makes the optimistic hypotheses that the number of sources is equal to the dimension of the signal (the number of captors), and that the unknown mixture matrix is invertible. The goal of BSS is then to compute an estimate of the inverse of the mixture matrix in order to extract from the signal the independent components.

In the present paper we study the effect of having sources with different "strengths" when performing BSS. After giving a proper definition of the strength of a source, the main purpose of our study is to relate the strength of a source to its contribution to the information conveyed by the processing system about the signal, and to consider in more detail the case where some of the sources are very weak compared to the others. We will show that in that case it is worthwhile to project the data onto the space generated by the strong sources in order to extract meaningful information and to avoid numerical problems. The contributions to the (projected) signal from the weak sources can then be considered as noise terms added to the linear mixture of strong sources. Since the sources are independent, this "noise" is thus independent of the "pure" signal (the part due to the strong sources).

The paper is organized as follows. In Section 2 we introduce the model and give a precise definition to the strength of a source. In Section 3 we compute Shannon information quantities from which we characterize how each source contributes to the information conveyed by the data and by the output of the processing network. We then discuss the case of a linear mixture of N independent sources with $N - m$ "weak" sources and m "strong" sources. The results of Section 3 show that in such a case it would be preferable to be able to work in the m-dimensional space spanned by the strong sources. We show in Section 4 that, with a good approximation, this is simply done by projecting the data onto the m largest principal components. As a result one can perform BSS in the m-dimensional space where one is dealing with an m-dimensional linear mixture corrupted by a weak input noise. In Section 5 we study, at first nontrivial order in the noise strength, the expected performance in the estimation of the m strong sources. Eventually in Section 6 we present numerical simulations.

* Corresponding author. Permanent address. G. Nadjakov Institute for Solid State Physics, Bulgarian Academy of Sciences, 1784 Sofia, Bulgaria. Tel.: +34-91-398-7126; fax: +34-91-398-6697.

E-mail address: elka@fisfun.uned.es (E. Korutcheva).

[2] Present address. NASA/Goddard Institute for Space Studies, 2880 Broadway, New York, NY 10025, USA.

[3] Laboratoire associé au CNRS (URA 1306), à l'ENS et aux Universités Paris 6 et Paris 7.

PII: S0893-6080(00)00041-1

2. The model

We consider the information processing of a signal which is an N-dimensional linear mixture of N independent sources. At each time t one observes $\mathbf{S}(t) = \{S_j(t),\ j = 1,\ldots,N\}$, which can be written in term of the unknown sources $\mathbf{s}(t) = \{s_\alpha(t),\ \alpha = 1,\ldots,N\}$ as:

$$S_j = \sum_{\alpha=1}^{N} M_{j\alpha} s_\alpha, \qquad j = 1,\ldots,N, \tag{1}$$

where $\mathbf{M} = \{M_{j\alpha},\ j = 1,\ldots,N,\ \alpha = 1,\ldots,N\}$ is the mixture matrix assumed to be invertible. As it is well known, and easily seen from the above equation, it is not possible to distinguish between the mixture of $\mathbf{s}$ with the matrix $\mathbf{M}$ from the mixture of $\mathbf{s}' \equiv \mathbf{PDs}$ with the matrix $\mathbf{M}' \equiv \mathbf{MD}^{-1}\mathbf{P}^{-1}$ where $\mathbf{D}$ is an arbitrary diagonal matrix with non-zero diagonal elements, and $\mathbf{P}$ an arbitrary permutation of N indices. If we decide to consider both normalized sources and normalized mixture matrices, we are left with a diagonal matrix $\mathbf{D}$, which defines the "strengths" of the sources. More precisely we write

$$S_j = \sum_{\alpha=1}^{N} \bar{M}_{j\alpha} \eta_\alpha s_\alpha, \qquad j = 1,\ldots,N \tag{2}$$

assuming zero mean and unit variance for every source:

$$\langle s_\alpha \rangle = 0, \qquad \langle s_\alpha^2 \rangle = 1, \qquad \alpha = 1,\ldots,N, \tag{3}$$

where $\langle\ \rangle$ denotes the average with respect to the (unknown) sources probability distributions,

$$\rho(\mathbf{s}) = \prod_\alpha \rho_\alpha(s_\alpha), \tag{4}$$

with $\bar{\mathbf{M}}$ the normalized mixture matrix. The normalization can be chosen in different ways, and the two of them are of particular interest for what follows. The simplest one is, for each α

$$[\bar{\mathbf{M}}^{\mathrm{T}}\bar{\mathbf{M}}]_{\alpha\alpha} = \sum_{j=1}^{N} (\bar{M}_{j\alpha})^2 = 1. \tag{5}$$

The second one is a normalization on the inverse of the mixture matrix:

$$[\bar{\mathbf{M}}^{-1}\bar{\mathbf{M}}^{\mathrm{T}-1}]_{\alpha\alpha} = \sum_{j=1}^{N} ([\bar{M}^{-1}]_{\alpha j})^2 = 1. \tag{6}$$

Once a particular normalization, such as Eqs. (5) or (6), is chosen, the parameters η_α in Eq. (2) are well defined and can be understood as the relative strengths of the sources.

3. Information processing in the presence of inhomogeneous sources

Since the mixture matrix is assumed to be invertible, it is in principle possible to compute an estimate of it. This can be done with any one of the known BSS algorithms (see, e.g. Bell & Sejnowki, 1995; Cardoso, 1989; Comon, 1994; Nadal & Parga, 1997). As a result one obtains an estimate of the inverse of the mixture matrix, which in our notations can be written as

$$\frac{1}{\eta_a}[\bar{M}^{-1}]_{a,j}. \tag{7}$$

This shows that it will be dominated by the smallest ηs, and numerical instabilities or overflows may occur if some of them are very small. In many approaches to BSS whitening of the data is first performed. The whitened data are then an orthogonal mixture of sources, so that after this preprocessing one has sources of equal strengths. But this preprocessing requires a multiplication by the inverse of the eigenvalues, and this is subject to the same numerical problems as with the computation of the inverse of the mixture matrix: as we will see in Section 4, small values of η lead to the existence of small eigenvalues.

3.1. Information content of the data

Let us now compute the amount of information conveyed by the data, $\mathbf{S}$, about the sources, that is the mutual information (Blahut, 1988) $I(\mathbf{S},\mathbf{s})$. To do so we consider

$$S_j = \sum_{\alpha=1}^{N} \bar{M}_{j\alpha} \eta_\alpha s_\alpha + \nu_j, \qquad j = 1,\ldots,N. \tag{8}$$

where $\boldsymbol{\nu} = \{\nu_j,\ j = 1,\ldots,N\}$ is a vanishing additive noise, $\langle \nu_j \rangle = 0$, $\langle \nu_j \nu_k \rangle = b\delta_{j,k}$ with $b \to 0$. Then $I(\mathbf{S},\mathbf{s})$ is a constant (that is a quantity that depends on b alone) plus the data entropy. Since the mixture matrix is invertible, we have

$$I(\mathbf{S},\mathbf{s}) = \text{Const.} + \ln|\det\bar{\mathbf{M}}| + \sum_\alpha \ln\ \eta_\alpha - \sum_\alpha \int \mathrm{d}\mathbf{h}_\alpha \rho_\alpha(\mathbf{h}_\alpha) \ln\ \rho_\alpha(\mathbf{h}_\alpha). \tag{9}$$

The last term in the above expression is the sum of the source entropies. One should remember that the s's are the normalized sources, $\langle s_\alpha^2 \rangle = 1$. This shows that each source contributes to the information by a combination of its strength and its entropy: the strength term favors strong sources, whereas the entropy term favors the sources with a probability distribution function (p.d.f.) close to Gaussian. The entropy terms, however, are bounded: the entropy of a source cannot exceeds the one of a Gaussian with same variance, that is

$$-\int \mathrm{d}h_\alpha \rho_\alpha(h_\alpha) \ln\ \rho_\alpha(h_\alpha) \leq \frac{1}{2} \ln\ 2\pi\mathrm{e}. \tag{10}$$

Hence the information can be easily dominated by the strength terms, which can be arbitrarily large.

It is known that for performing BSS perfect knowledge of the sources distribution is not necessary and that working on the cumulants of order 2 and 3 or 4 is sufficient (see, e.g.

Comon, 1994; Nadal & Parga, 1997). We can thus analyze the result, Eq. (9), by making a close-to-Gaussian approximation (Comon, 1994; Nadal & Parga, 1997). If we assume the sources to have non-zero third-order cumulants,

$$\lambda_\alpha^{(3)} \equiv \langle s_\alpha^3 \rangle_c, \tag{11}$$

we replace the source distribution ρ_α by

$$\hat{\rho}_\alpha(s_\alpha) = \frac{e^{-s_\alpha^2/2}}{\sqrt{2\pi}}\left(1 + \lambda_\alpha^{(3)}\frac{s_\alpha(s_\alpha^2-3)}{6}\right). \tag{12}$$

The distribution $\hat{\rho}_\alpha$ has the same three first moments as the true distribution ρ_α (Abramowitz & Stegun, 1972).

In the case of a symmetric non-Gaussian distribution, the third-order cumulants are zero and one has then to take into account non-zero fourth-order cumulants. It is a straightforward exercise to perform the same analysis as below in that case. For simplicity in this paper we will consider only the case of non-symmetric distributions.

Within this approximation, Eq. (12), the mutual information (9) reads:

$$I(\mathbf{S},\mathbf{s}) = \text{Const.} + \ln|\det\bar{\mathbf{M}}| + \sum_\alpha \ln\,\eta_\alpha + \frac{N}{2}\ln 2\pi e - \frac{1}{12}\sum_\alpha \langle s_\alpha^3\rangle_c^2. \tag{13}$$

From the above expression, the most important sources are those for which the quantity

$$\langle s_\alpha^3\rangle_c^2 - \ln\,\eta_\alpha \tag{14}$$

is the smallest.

We consider now the information that will be conveyed by a network processing the data, and ask for the contribution to this information by each source when the network performs BSS.

3.2. Characterization from infomax

The infomax criterion (Linsker, 1988; Nadal & Parga, 1994) will allow us to get some more insight into the link between the sources strengths and the amount of information that can be extracted from the data.

We consider the information processing of the signal by a non-linear network, and we are interested in computing the mutual information $I(\mathbf{V},\mathbf{S})$ between the input $\mathbf{S}$ and the output $\mathbf{V} = \{V_i,\ i = 1,\ldots,N\}$ of the network. Since the signal is a linear mixture, the relevant architecture is a linear processing followed by a (possibly) nonlinear transfer function that may differ from neuron to neuron:

$$V_i = f_i(h_i) + \nu_i \tag{15}$$

$$h_i = \sum_j J_{ij}(S_j + \nu_j^0), \tag{16}$$

where $\boldsymbol{\nu}_0 = \{\nu_j^0,\ j = 1,\ldots,N\}$ and $\boldsymbol{\nu} = \{\nu_i,\ i = 1,\ldots,N\}$ are additive input and output noises, respectively, with $\langle\boldsymbol{\nu}_0\rangle = 0$, $\langle\boldsymbol{\nu}\rangle = 0$, $\langle\nu_j^0\nu_{j'}^0\rangle = b^0\delta_{j,j'}$, $\langle\nu_i\nu_{i'}\rangle = b\delta_{i,i'}$.The J_{ij} can be viewed as synaptic efficacies and the h_is as post-synaptic potentials (PSP). As explained in the previous section, the noise has to be introduced in order to have a non-trivial mutual information, and we take the limit $0 \leq b^0 \ll b \ll 1$. For strictly zero input noise, $b^0 = 0$, in the limit $b \to 0$ the mutual information is up to a constant equal to the output entropy. As shown in Nadal and Parga, (1994) its maximization over the choice of both $\mathbf{J}$ and the transfer functions f_i's leads to BSS. One can then derive practical algorithms for performing BSS (Bell & Sejnowski, 1995). In this limit of $b^0 = 0$ all the sources play the same role, that is the maximum of the mutual information is independent of the individual sources properties as well as of the mixture matrix. When one takes into account a non-zero input noise, then at first non-trivial order in b^0/b one sees that the input noise introduces a scale that breaks this invariance. More precisely, at first-order in b^0/b the mutual information $I(\mathbf{V},\mathbf{S})$ can be written (see Nadal and Parga (1994) for details):

$$I(\mathbf{V},\mathbf{S}) = I_0(\mathbf{V},\mathbf{S}) - \frac{b^0}{2b}\sum_{i=1}^N \Gamma_{ii}\int dh_i\psi_i(h_i)f_i'^2, \tag{17}$$

where $I_0(\mathbf{V},\mathbf{S})$ is the value at $b^0 = 0$,

$$I_0(\mathbf{V},\mathbf{S}) = \text{Const.} - \int d\mathbf{h}\psi(\mathbf{h})\ln\frac{\psi(\mathbf{h})}{\prod_{i=1}^N f_i'(h_i)} \tag{18}$$

and $(b^0/b)\Gamma_{ii}$ is the variance of the noise on the PSP h_i:

$$\Gamma_{ii} \equiv [\mathbf{J}\mathbf{J}^{\mathrm{T}}]_{ii}. \tag{19}$$

Finally, $\psi(\mathbf{h})$ is the probability distribution of $\mathbf{h}$ induced by the sources input distribution, and $\psi_i(h_i)$ the marginal distribution of the PSP h_i. At a given $\mathbf{J}$, optimizing with respect to the choice of transfer functions gives

$$f_i'(h_i) = \psi_i(h_i)\left\{1 + \frac{b^0}{b}\Gamma_{ii}[\langle\psi_i^2\rangle - \psi_i^2(h_i)]\right\} \tag{20}$$

with $\langle\psi_i^2\rangle = \int dh_i\psi_i(h_i)\psi_i^2(h_i) = \int dh_i\psi_i(h_i)^3$. We now optimize over $\mathbf{J}$. At zeroth-order the optimum is reached for $\mathbf{J} = \mathbf{M}^{-1}$ (up to an arbitrary permutation), so that we write

$$\mathbf{W} \equiv \mathbf{J}\mathbf{M} = \mathbf{1}_N + \frac{b^0}{b}\mathbf{W}^1, \tag{21}$$

where $\mathbf{1_N}$ is the $N \times N$ identity matrix. Expanding the mutual information at first-order in b^0/b one finds that there is no contribution from $\mathbf{W}^1$ to this order. Hence the mutual information at first-order in b^0/b is given by Eq. (17) at $\mathbf{J} = \mathbf{M}^{-1}$, with f_i' given by Eq. (20) in which we set $\psi_i = \rho_i$. This gives

$$I(\mathbf{V},\mathbf{S}) = \text{Const.} - \frac{b^0}{2b}\sum_{\alpha=1}^N \Gamma_{\alpha\alpha}\int ds_\alpha[\rho_\alpha(s_\alpha)]^3 \tag{22}$$

with

$$\Gamma_{\alpha\alpha} = [\mathbf{M}^{-1}\mathbf{M}^{\mathrm{T}-1}]_{\alpha\alpha}. \tag{23}$$

One sees that the term depending on $\mathbf{M}$ is what appears in normalization (6) of the mixture matrix. Hence if one chooses this particular normalization (6) in order to define the strengths, η_α, of the sources, one can rewrite

$$I(\mathbf{V},\mathbf{S}) = \text{Const.} - \frac{b^0}{2b}\sum_{\alpha=1}^{N}\frac{1}{\eta_\alpha^2}\langle\rho_\alpha^2\rangle \tag{24}$$

with $\langle\rho_\alpha^2\rangle = \int \mathrm{d}s_\alpha[\rho_\alpha(s_\alpha)]^3$. The above expression shows how each source contributes to the mutual information in term of its strength η_α and its p.d.f. ρ_α.

Within the close-to-Gaussian approximation (12) one gets

$$I(\mathbf{V},\mathbf{S}) = \text{Const.} - \frac{b^0}{b}\sum_{\alpha=1}^{N}\langle s_\alpha^3\rangle_c^2\frac{1}{\eta_\alpha^2}. \tag{25}$$

Hence the sources that contribute the most to the conveyed information are those for which the quantity

$$\mathscr{E}_\alpha \equiv \langle s_\alpha^3\rangle_c^2\frac{1}{\eta_\alpha^2} \tag{26}$$

is the smallest. One should remember that, here, η_α is given by

$$\frac{1}{\eta_\alpha^2} = \sum_{j=1}^{N}([\mathbf{M}^{-1}]_{\alpha j})^2. \tag{27}$$

3.3. Discussion

As already seen when computing the mutual information between the data and the sources, a source will contribute if it is strong and/or close to Gaussian. However, the particular combination that appears here is different from the one we obtained in the previous section: here we have a multiplicative combination of strength and cumulant, whereas in Eq. (14) it was an additive combination.

An important practical remark is that, if the third-order cumulants are zero, the close-to-Gaussian approximation has to take into account the fourth-order cumulants. Then, instead of Eqs. (14) and (26) one gets similar expressions with the fourth-order cumulants in place of the third-order ones.

The criterion (26) can be used in different ways, depending on the particular application considered. The quantity $\mathscr{E}_\alpha$ is zero for Gaussian sources, whatever their strengths. This is not surprising since the Shannon information is maximal for Gaussian distributions. However, in many cases the Gaussian part of the signal is considered as "noise", and the non-Gaussian part is the "meaningful" part, the "true" signal. Hence mutual information can be used as a cost function in order to extract this noise, in particular when it is strong, which can then be subtracted from the input signal. In cases where one has distributions of similar shapes, Eq. (26) suggests to use the strength as defined in Eq. (27) to order the sources and select the most relevant ones.

To conclude this section, we see that the intuitive idea that weak sources can be considered as noise terms and cannot be estimated, can be quantified from various point of views. From the purely numerical aspect, the mixture matrix is close to being singular; the information content of the data, the amount of information conveyed by a processing channel, are seriously diminished by the presence of weak sources. From this analysis, it appears clearly that it would be preferable to be able to project the data onto the space spanned by the strong sources, in order to work in a space of smaller dimension with sources of similar strengths. In the next section we show that this is simply done by making use of the principal component analysis.

4. Principal component analysis

A standard approach in data processing consists in first performing the principal component analysis (PCA), and then projecting the data onto the eigenspace associated with the largest eigenvalues. In the present context of BSS, it is reasonable to expect the space spanned by the strong sources to be essentially the same as the one associated to the largest principal components. It is the purpose of this section to give a positive and more precise answer to this question.

We consider the specific case where m sources are strong, while $N-m$ sources are weak. More precisely, choosing for later convenience normalization (5), we assume

$$\begin{aligned}&\eta_\alpha \sim O(1=\epsilon^0) \ \text{ for } \alpha = 1,\ldots,m\\ &\eta_\alpha \sim O(\epsilon) \ \text{ for } \alpha = m+1,\ldots,N,\end{aligned} \tag{28}$$

where ϵ is a small parameter, $\epsilon \ll 1$. This is equivalent to state that there is a gap in the spectrum of eigenvalues at the λ_m, with $\lambda_{m+1} \ll \lambda_m$.

We assume that the reduced $N\times m$ mixture matrix M^0, $\{M^0_{j\alpha} = M_{j\alpha},\ j=1,\ldots,N;\ \alpha=1,\ldots,m\}$ is of rank m, so that the $(N\times N)$ correlation matrix (the covariance of the input signal) $\mathbf{C}^0$, which would be obtained at $\epsilon \equiv 0$, has m non-zero eigenvalues. It is a standard exercise in perturbation theory (Messiah, 1961) to study the behavior of the eigenvalues and eigenvectors of a symmetric matrix, here the covariance matrix $\mathbf{C}$ of the inputs, at first non-trivial order in the small parameter ϵ. The eigenvalues have a smooth behavior with ϵ: the m largest eigenvalues of $\mathbf{C}$ are, at first non-trivial order, the m non-zero eigenvalues of $\mathbf{C}^0$ shifted by quantities of order ϵ^2, and the $N-m$ smallest ones are of order ϵ^2. However, the eigenvectors are very sensitive to small variations of ϵ–this is related to the fact that the mixture matrix M is closed to be singular for small ϵ. More precisely, one gets the following results.

One can write $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{C}^0 + \epsilon^2 \mathbf{C}^1, \tag{29}$$

where $\mathbf{C}^0$ is the correlation of the inputs that would be obtained without the weak sources ($\epsilon \equiv 0$), and $\epsilon^2\mathbf{C}^1$ contains all the contributions from the weak sources. We denote by λ_α^0 the eigenvalues of $\mathbf{C}^0$, with $\{\lambda_\alpha^0,\ \alpha = 1,\ldots,m\}$ non-zero and $\lambda_\alpha^0 = 0$ for $\alpha = m+1,\ldots,N$. The associated eigenvectors $\{\mathbf{v}_\alpha^0,\ \alpha = 1,\ldots,N\}$ form an orthonormal basis. If all the eigenvalues of $\mathbf{C}^0$ are different (hence in particular $N = m+1$), then, at first order, the eigenvalues of $\mathbf{C}$ are

$$\lambda_\alpha = \lambda_\alpha^0 + \epsilon^2 \lambda_\alpha^1$$

$$\lambda_\alpha^1 = \mathbf{v}_\alpha^{0\mathrm{T}} \mathbf{C}^1 \mathbf{v}_\alpha^0 \qquad (\alpha = 1,\ldots,N), \tag{30}$$

and the corresponding eigenvectors are

$$\mathbf{v}_\alpha = \mathbf{v}_\alpha^0 + \epsilon^2 \sum_{\beta \neq \alpha} \mathbf{v}_\beta^0 \frac{\mathbf{v}_\alpha^{0\mathrm{T}} \mathbf{C}^1 \mathbf{v}_\beta^0}{\lambda_\alpha^0 - \lambda_\beta^0} \qquad (\alpha = 1,\ldots,N). \tag{31}$$

If there are degenerate eigenvalues (in particular the null eigenvalue is degenerate for $N > m+1$), this is modified as follows. Suppose $\mathbf{C}^0$ has only $r < N$ different eigenvalues, $\mu_1 > \mu_2 > \cdots > \mu_r$, with degeneracies q_a, $a = 1,\ldots,r$ ($\sum_a q_a = N$, $\mu_r = 0$ if $N > m+1$). We have

$$\lambda_\alpha^0 = \mu_a \qquad \text{for } \sum_{b=1}^{a-1} q_b < \alpha \leq \sum_{b=1}^{a} q_b \equiv \alpha_a \tag{32}$$

and we set $\alpha_0 \equiv 0$. Consider an eigenvalue μ_a with degeneracy $q_a > 1$. The eigenvectors of $\mathbf{C}^0$ associated to μ_a, $\{\mathbf{v}_\alpha^0,\ \alpha_{a-1} < \alpha \leq \alpha_a\}$, form an orthonormal basis of this eigenspace of dimension q_a, and this base is defined up to an arbitrary orthogonal transformation. This arbitrariness is removed at first non-trivial order in ϵ, together with the removal of the eigenvalue degeneracy: the new q_a eigenvalues for $\{\alpha_{a-1} < \alpha \leq \alpha_a\}$ are given by Eq. (30), where the $\mathbf{v}^0$s form the particular $q_a \times q_a$ orthogonal matrix that diagonalizes $\mathbf{C}_a^1$, the restriction of the matrix $\mathbf{C}^1$ to the eigenspace of μ_a, λ_α^1 being then the eigenvalue of $\mathbf{C}_a^1$.

The eigenvectors $\mathbf{v}$ are now given by an equation similar to Eq. (31), with the sum over $\beta \neq \alpha$ replaced by a sum over the β such that $\lambda_\beta \neq \lambda_\alpha$, and a new term specific to each degenerate eigenvalue μ_a:

$$\mathbf{v}_\alpha = \mathbf{v}_\alpha^0 + \epsilon^2 \sum_{\beta:\lambda_\beta \neq \lambda_\alpha} \mathbf{v}_\beta^0 \frac{\mathbf{v}_\alpha^{0\mathrm{T}} \mathbf{C}^1 \mathbf{v}_\beta^0}{\lambda_\alpha^0 - \lambda_\beta^0} + \epsilon^2 \sum_{\beta:\lambda_\beta = \lambda_\alpha} X_{\alpha,\beta} \mathbf{v}_\beta^0 \tag{33}$$

$(\alpha = 1,\ldots,N)$,

where the $\mathbf{v}^0$ are chosen as just explained, and $X_{\alpha,\beta}$ is an arbitrary antisymmetric matrix.

The final result is thus that the space generated by the m eigenvectors associated to the m largest eigenvalues is, to order ϵ^2, the same space as the one that would be obtained in the absence of the weak sources. Projecting the data onto this space is then equivalent to working with the m-dimensional signal which is the mixture of the m strong sources, weakly corrupted by an additive noise.

5. BSS with noisy data

Let us now assume that we have preprocessed the data by projecting it onto the m largest principal components. To avoid the introduction of a new notation, in the following $\{S_j,\ j = 1,\ldots,m\}$ will denote these preprocessed data (projections) instead of the data themselves. Instead of the model Eq. (1) we have thus to consider the model

$$S_j = \sum_{\alpha=1}^{m} M_{j\alpha} s_\alpha + \nu_j^0, \qquad j = 1,\ldots,m. \tag{34}$$

The matrix $\mathbf{M}$ is now a $m \times m$ invertible mixture matrix, such that $\mathbf{M}\mathbf{M}^{\mathrm{T}}$ has m non-zero, of order $1 = \epsilon^0$, eigenvalues. The s_α's ($\alpha = 1,\ldots,m$) are the sources of interest, and the ν_j^0's are additive noises, resulting from the weak sources, as explained in the previous section. This noise $\boldsymbol{\nu}_0 = \{\nu_j^0,\ j = 1,\ldots,m\}$ is uncorrelated with the m (strong) sources, and of arbitrary distribution $P(\boldsymbol{\nu}_0)$. Since we are working in the small ϵ regime, all we will need is to characterize this distribution by its first two cumulants:

$$\langle \boldsymbol{\nu}_0 \rangle = 0$$

$$\langle \nu_0 \nu_0^{\mathrm{T}} \rangle = \epsilon^2 \mathbf{B}, \tag{35}$$

where $\mathbf{B}$ is a (possibly non-diagonal) $m \times m$ symmetric matrix. The problem we are considering now is thus strictly the same as the one of performing BSS on a linear mixture of m sources corrupted by some additive input noise, which, although small, cannot be neglected.

5.1. The mutual information

In this section we consider this noisy BSS problem within the infomax approach as formulated in Nadal and Parga (1994). The network we consider has the same architecture as the one defined in Eq. (16), but with m inputs and outputs:

$$V_i = f_i(h_i) + \nu_i \tag{36}$$

$$h_i = \sum_{j=1}^{m} J_{ij}(S_j + \nu_j^0) \qquad i = 1,\ldots,m, \tag{37}$$

with $\langle \nu_i \nu_{i'} \rangle = b\delta_{i,i'}$. The limit to be considered here is the one of a vanishing output noise, $b \to 0$, but at a given input noise level:

$$0 < b \ll \epsilon^2. \tag{38}$$

Another important difference with the calculation done in Section 3.2, is that here we are interested in computing the information conveyed about the global input, $\mathbf{S} + \boldsymbol{\nu}_0$, and not about the "pure" signal alone, $\mathbf{S}$. Indeed, in Section 3.2 we considered some input noise corresponding to some noise at the level of the receptors, whereas here the actual

signal is the global input, $\mathbf{S} + \boldsymbol{\nu}_0$, in which *we* have decided to call "(pure) signal" the part coming from the strong sources and "noise" the part due to the weak sources.

In this limit of vanishing output noise, the mutual information $I(\mathbf{V}, \mathbf{S} + \boldsymbol{\nu}_0)$ between the output and the input of the network is up to a constant equal to the output entropy. To simplify the analysis, we assume a full adaptation of the transfer functions, which means (Nadal & Parga, 1994), for $\mathbf{J}$ given,

$$f_i'(h_i) = \psi_i(h_i), \qquad i = 1, \ldots, m, \tag{39}$$

where $\psi_i(h_i)$ is the marginal probability distribution of the PSP h_i. As a result the mutual information is up to a constant equal to the redundancy between the PSPs (Nadal & Parga, 1994):

$$I(\mathbf{V}, \mathbf{S}) = \text{Const.} - \int \mathrm{d}^m \mathbf{h} \psi(\mathbf{h}) \ln \frac{\psi(\mathbf{h})}{\prod_{i=1}^{m} \psi_i(h_i)}. \tag{40}$$

5.2. Maximization in the small ϵ limit

In term of the sources distributions, the distribution $\psi(\mathbf{h})$ is given by:

$$\psi(\mathbf{h}) = \int \prod_{\alpha=1}^{m} \mathrm{d}s_\alpha \rho_a(s_\alpha) \times \int \mathrm{d}^m \boldsymbol{\nu}_0 P(\boldsymbol{\nu}_0) \prod_{i=1}^{m} \delta\left(h_i - \sum_\alpha [JM]_{i\alpha} s_\alpha - \sum_j J_{ij} \nu_j^0 \right). \tag{41}$$

Since in Eq. (41) the noises ν_j^0 are $\sim O(\epsilon)$ we can perform an expansion, leading to the following expression:

$$\psi(\mathbf{h}) = \left\{ 1 + \frac{\epsilon^2}{2} \sum_{i,i'} [\mathbf{J}\mathbf{B}\mathbf{J}^{\mathrm{T}}]_{ii'} \partial_i \partial_{i'} \right\} \psi^0(\mathbf{h}), \tag{42}$$

where ∂_i means the partial derivative with respect to h_i, and $\psi^0(\mathbf{h})$ is the p.d.f. that would be obtained at $\epsilon = 0$. Because the noise has zero mean there is no term of order ϵ in Eq. (42).

We consider now the maximization of the mutual information over the choice of $\mathbf{J}$, taking into account that ϵ is small. If ϵ was strictly zero, we would be back to the noiseless BSS problem for which the optimum is reached for $\mathbf{J} = \mathbf{M}^{-1}$ (up to an arbitrary permutation). So for non-zero ϵ we write

$$\mathbf{W} \equiv \mathbf{J}\mathbf{M} = \mathbf{1}_m + \epsilon \mathbf{W}^1 + O(\epsilon^2), \tag{43}$$

where $\mathbf{1}_m$ is the $m \times m$ identity matrix, and the correction is a matrix of order at least ϵ. Since $\mathbf{W}$ depends now on ϵ we can also expand ψ^0 in powers of ϵ, and finally $\psi(\mathbf{h})$ can then be written as

$$\psi(\mathbf{h}) = \left[\prod_\alpha \rho_\alpha(h_\alpha) \right] [1 + \epsilon Q[\mathbf{h}] + R[\mathbf{h}]] \tag{44}$$

with

$$Q[\mathbf{h}] \equiv - \sum_{\alpha,\beta} [\ln \rho_\alpha]' W^1_{\alpha\beta} h_\beta - \mathrm{Tr}\mathbf{W}^1 \tag{45}$$

and $R[\mathbf{h}]$ contains terms of order at least ϵ^2, coming from both $\mathbf{W}$, Eq. (43), and $\mathbf{B}$, Eq. (42). Similarly, for the marginal distributions:

$$\psi_\alpha(h_\alpha) = \rho_\alpha(h_\alpha) \{ 1 + \epsilon Q_\alpha[h_\alpha] + R_\alpha[h_\alpha] \}, \tag{46}$$

with

$$Q_\alpha[h_\alpha] \equiv -[\ln \rho_\alpha]' W^1_{a\alpha} h_\alpha - W^1_{\alpha\alpha}. \tag{47}$$

The substitution of Eqs. (44) and (46) in expression (40) gives then for the mutual information, at first non-trivial order:

$$I(\mathbf{V}, \mathbf{S}) = I_0(\mathbf{V}, \mathbf{S}) - \frac{\epsilon^2}{2} \int \prod_{\alpha=1}^{m} \mathrm{d}h_\alpha \rho_\alpha(h_\alpha) \times \left[Q[\mathbf{h}] - \sum_\alpha Q_\alpha[h_\alpha] \right]^2. \tag{48}$$

The term $I_0(\mathbf{V},\mathbf{S})$ corresponds to the part of the mutual information that does not take into account the weak sources. It is the same as if one computes the mutual information between the output $\mathbf{V}$ and the signal $\mathbf{Ms}$; $I(\mathbf{V},\mathbf{Ms})$. The fact that there is no term of order ϵ in Eq. (48) can be understood as coming from the normalization conditions $\int \mathrm{d}\mathbf{h} \psi^0(\mathbf{h}) = 1$ and $\int \mathrm{d}h_\alpha \psi^0_\alpha = 1$, which imply

$$\int \prod_{\alpha=1}^{m} \mathrm{d}h_\alpha \rho_\alpha(h_\alpha) Q[\mathbf{h}] = 0$$

and

$$\int \mathrm{d}h_\alpha \rho_\alpha(h_\alpha) Q_\alpha[h_\alpha] = 0$$

(these properties can be easily checked by performing the integrations using the explicit expressions (45) and (47)). One has similar properties for the quantities of order ϵ^2, $R[\mathbf{h}]$ and $R_\alpha[h_\alpha]$ defined in Eqs. (44) and (46), so that they do not contribute at this order ϵ^2 in the final result (48).

Now one has

$$Q[\mathbf{h}] - \sum_\alpha Q_\alpha[h_\alpha] = - \sum_{\alpha \neq \beta} [\ln \rho_\alpha]' W^1_{\alpha\beta} h_\beta. \tag{49}$$

The mutual information is maximized when the quadratic term in Eq. (48) is minimized, that is for $W^1_{\alpha\beta} = 0$ for $\alpha \neq \beta$. It follows that there is no correction to the mutual information at order ϵ^2 and that corrections due to the weak sources appear at order ϵ^4.

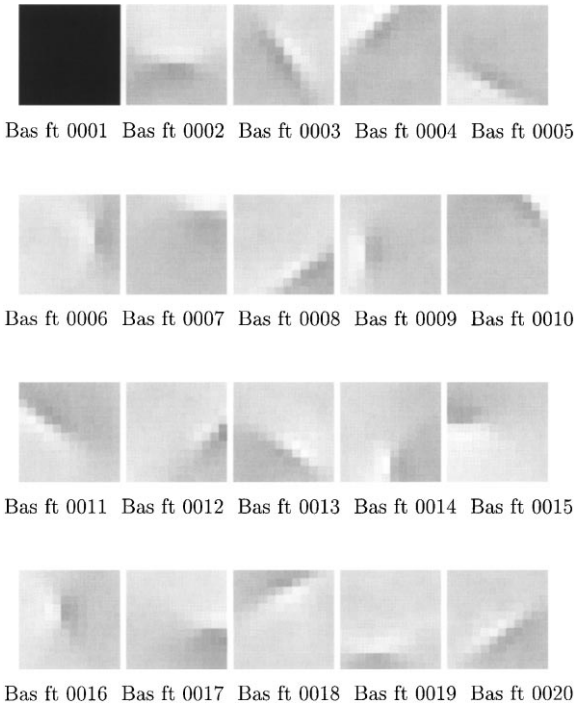

Fig. 1. Basis functions of the ICA solution.

6. Numerical simulations

In this section we illustrate our analysis by numerical simulations. We test the above analysis on the following toy example. We consider the ICA of natural images performed in Bell & Sejnowski (1995). First we reproduce the results in Bell and Sejnowski (1995) (not shown here). We then create a new database with artificially increased component strengths: new images are computed as a linear mixture of the previous ICA basis function but the strength of 20 components was augmented 100 times compared to the other 124. We performed ICA in this new data base, with the same algorithm based on infomax (Bell & Sejnowski, 1995; Nadal & Parga, 1994), but after projecting the data onto the 20 largest principal components. The resulting basis function represented in Fig. 1 shows the efficiency of PCA preprocessing: we find the good 20 stronger components and the computational time is considerably decreased.

For such a signal, the PCA analysis is identical to a Fourier analysis, and therefore dropping the smallest eigenvalues means neglecting high frequencies. One thus expects to extract components that are smoothed versions of components extracted when working with the full space. This is indeed the case as shown in Fig. 1.

7. Concluding remarks

We have discussed the task of BSS in the case of a mixture of sources of unequal strengths. We have presented different, but related, ways of defining the relative strengths of the sources. In particular, when non-zero input noise is taken into account the contribution of a source to the conveyed information can be characterized by a criterion that combines the mixture matrix elements and the third cumulant of the source distribution. This allows to define the strength of a source once a proper normalization of the mixture matrix is assumed. Conversely, this study shows which sources will be “preferred” by the infomax criterion (which part of the signal is more likely to be well extracted by an ICA performed with infomax).

The analysis indicates also that, although arbitrary, the assumed normalization of the mixture matrix may have an important practical role in the analysis of the outcome of an ICA, whenever one wants to extract the “meaningful” sources. Which part of the signal is more important is of course an application-dependent notion. Prior knowledge related to a given case should allow to define the proper normalization from which the appropriate scale of source strengths can be defined. Conversely each chosen normalization implies a particular physical interpretation that should be kept in mind when analyzing the outcome of an ICA.

We have considered, in more detail, the particular case of the information processing of a linear mixture of independent sources when some of them are very weak as compared to the other sources. One should note that in such case the notion of strong versus weak is independent of the mixture matrix normalization. It is easily seen that the presence of weak sources leads to an almost singular mixture matrix, and this manifests itself by the existence of very small eigenvalues in the PCA analysis. We have shown that it is relevant to project the input data onto the largest principal components in order to extract the strongest independent sources. We have thus quantified the intuitive idea that the subspace, where most of the data live, is mainly spanned by the strongest independent sources. We illustrated this result on the ICA of the image data base studied in Bell & Sejnowski (1995).

A possible situation where the PCA will not be (sufficiently) helpful is when the strong sources generate a linear space of dimension smaller than the number of sources. This space will be found by the PCA. After projection onto the largest PCs, one has then to deal with an ICA with a number of sources larger than the number of captors. This is an interesting problem that has received considerable attention recently, and several algorithms have been proposed. Our analysis suggests then that it can be meaningfull to project onto the largest PCs (in order to eliminate the weak sources) and yet to search for a number of (strong) ICs larger than the number of largest PCs.

Acknowledgements

We thank T. Bell for giving us access to his image database. This work has been partly supported by the French

contract DGA 962557A/DSP. E.K. warmly thanks for hospitality the Laboratoire de Physique Statistique de l'Ecole Normale Superieure, where this work was performed. E.K. has been also supported by the Spanish DGES project PB97-0076 and partly by the French–Spanish Program PICASSO and the Bulgarian Scientific Foundation Grant F-608.

References

Abramowitz, M., & Stegun, I. A. (1972). *Handbook of mathematical functions*, New York: Dover.

Amari, S.-I., & Cardoso, J. -F. (1997). Blind source separation-semiparametric statistical approach. *IEEE Transactions on Signal Processing, Special issue on neural networks*, *45* (11), 2692.

Bell, A., & Sejnowski, T. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*, 1129–1159.

Blahut, R. E. (1988). *Principles and practice of information theory*, Cambridge, MA: Addison-Wesley.

Cardoso, J.-F. (1989). Source separation using higher-order moments. In *Proceedings of the International Conference on Acoustics, Speech Signal Process*, Glasgow (pp. 2109–2112).

Comon, P. (1994). Independent component analysis, a new concept?. *Signal Processing*, *36*, 287–314.

Herault, J., Jutten, C., & Ans, B. (1985). Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In *Proceedings of GRETSI*, Nice (pp. 1017–1020).

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*, 105–117.

Messiah, A. (1961). *Quantum mechanics*, Amsterdam: Elsevier.

Nadal, J.-P., & Parga, N. (1994). Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network: Computation in Neural Systems*, *5*, 565–581.

Nadal, J.-P., & Parga, N. (1997). Redundancy reduction and independent component analysis: algebraic and adaptive approaches. *Neural Computation*, *9* (7), 1421–1456.